Brain Tumor Segmentation Meets Efficiency: Res-UNet Improved by Attention Mechanisms and Quantization

Kasiful Aprianto[1] and Dwiza Riana[2*]
[1]Master of Computer Science, Faculty of Information Technology, Universitas Nusa Mandiri Jakarta, Indonesia 13620
[2]Department of Computer Science, Faculty of Information Technology, Universitas Nusa Mandiri Jakarta, Indonesia 13620
Email: [1]apriantokasiful@gmail.com, [2]dwiza@nusamandiri.ac.id

***Abstract*—Brain tumor segmentation from Magnetic Resonance Imaging (MRI) images is a crucial step in medical diagnosis and treatment planning, which directly impacts clinical decision-making and patient outcomes, particularly in resource-constrained medical environments. However, achieving high segmentation accuracy while maintaining computational efficiency remains a challenge, particularly for complex tumor types. Therefore, the research aims to use the brain tumor segmentation dataset and the brain tumor MRI dataset from Kaggle to evaluate segmentation performance. The analysis also investigates the trade-off between model accuracy and efficiency by optimizing the Res-UNet architecture with attention mechanisms, including the Attention Gate (AG), Squeeze-and-Excitation (SE) Block, and the Convolutional Block Attention Module (CBAM). As the result, attention mechanisms improve feature representation and segmentation precision. Then, these procedures also add computational cost. To address this challenge, Dynamic Range Quantization (DRQ) compresses the model from 127 MB to 32 MB (75% reduction) and speeds up inference by 37% (0.3143 s to 0.1973 s). During the process, the best model, Res-UNet with AG, achieves a mean Intersection over Union (IoU) of 0.845 and drops only by less than 0.0004 after quantization. Unlike previous studies that explored attention or quantization in isolation, the researchers combine both to achieve accurate, efficient, and deployable brain tumor segmentation for resource-constrained settings.**

***Index Terms*—Brain Tumor Segmentation, Res-UNet, Attention Mechanisms, Quantization**

I. Introduction

THE application of medical image segmentation is paramount in disease detection and diagnosis, particularly in the context of identifying brain tumors through Magnetic Resonance Imaging (MRI). Precise segmentation of brain tumors is crucial for various critical clinical tasks, including accurate diagnosis, individualized treatment planning (e.g., radiotherapy and surgical navigation), as well as effective monitoring of disease progression and treatment response. Moreover, inaccurate or inefficient segmentation can lead to misdiagnosis, suboptimal therapeutic interventions, and adverse patient outcomes. Deep learning methods, specifically Convolutional Neural Networks (CNNs), have shown outstanding performance in this domain due to the ability to extract complex features. Among these methods, U-Net has become a widely used architecture for medical image segmentation due to its ability to preserve spatial details across resolutions [1]. It has been successfully applied to various tasks [2]. To improve the performance of the U-Net for tumor detection, incorporating it with residual networks such as ResNet offers improved stability when training deeper models [3]. The resulting Res-UNet architecture enables deeper feature representation, improving segmentation accuracy for complex images, such as brain tumors.

Attention mechanisms, including the Attention Gate (AG), Squeeze and Excitation (SE) Block, and Convolutional Block Attention Module (CBAM), further enhance segmentation performance by highlighting relevant features while suppressing irrelevant ones. AG has proven effective in segmenting small organs such as the pancreas [4]. On the other hand, SE Block and CBAM strengthen major features as well as improve segmentation accuracy [5, 6]. Incorporating these mechanisms into Res-UNet is expected to improve segmentation metrics, such as Intersection over Union (IoU), a critical indicator of accuracy [7, 8].

Recent analysis explores compression methods to improve computational efficiency in this research. Quantization reduces model size while maintaining

Received: March 08, 2025; received in revised form: Aug. 15, 2025; accepted: Aug. 19, 2025; available online: Oct. 13, 2025.
*Corresponding Author

performance [9–11], and hybrid methods combining pruning as well as low-bit quantization achieve high accuracy with reduced computational cost [12, 13]. Adaptive quantization frameworks, such as Q-Net Compressor, have shown significant memory and power savings on constrained devices without compromising performance [13], making the models suitable for medical or mobile edge environments.

Many previous studies have investigated attention mechanisms to improve accuracy [4–8], or quantization to reduce complexity [9–13] in isolation, without systematically analyzing how the combination affects both segmentation accuracy and efficiency. This lack of incorporated evaluation limits practical deployment, specifically in resource-constrained clinical environments where both high accuracy and low computational cost are essential. To address this gap, the researchers integrate multiple attention mechanisms into a Res-UNet framework while simultaneously applying quantization techniques to compress the model. The proposed approach aims to balance segmentation accuracy with computational efficiency, providing a more comprehensive evaluation than prior isolated methods. Such integration is expected to yield a lightweight yet accurate segmentation model that is feasible for real-world medical applications, including mobile or embedded systems.

The research directly addresses the gap by systematically combining multiple attention mechanisms within the Res-UNet framework and applying Dynamic Range Quantization (DRQ) to achieve both high segmentation accuracy and computational efficiency. By developing a model that achieves an IoU of 0.845 while reducing model size by 75% and improving inference speed by 37%, the analysis provides a practical solution for brain tumor segmentation deployable on standard hospital workstations or edge devices. The main contributions are as follows: systematic investigation of various attention mechanisms (AG, SE Block, and CBAM) incorporated into the Res-UNet architecture for improved brain tumor segmentation accuracy; DRQ applications to achieve significant model size reduction and inference speed-up while preserving high segmentation performance; and comprehensive evaluation showing the optimal combination of attention mechanism (AG) and quantization.

A. Trade-off Between Attention Mechanisms and Quantization

Deep learning has revolutionized medical image segmentation, particularly in tasks requiring high precision, such as brain tumor segmentation. Attention mechanisms, including AG [4, 14], SE Block [5, 15], and CBAM [6, 16, 17], have been shown to improve segmentation accuracy by dynamically focusing on relevant regions while suppressing irrelevant information. This capability is particularly useful for segmenting complex structures such as brain tumors, where irregular shapes, subtle intensity variations, and unclear boundaries often pose significant challenges. However, as attention mechanisms generally improve feature extraction and segmentation accuracy, the improvement comes at a considerable cost. The incorporation can lead to increased model complexity, substantial computational demand, and higher memory usage [18–20]. Concerning clinical settings, specifically those in resource-limited regions or reliant on edge devices, such computational overhead renders these advanced models impractical due to excessive inference times, energy consumption, and stringent hardware requirements, limiting the real-world applicability.

Quantization addresses computational challenges by reducing the precision of model parameters, including weights and activations, from high-precision formats, such as FP32, to lower-precision formats, like INT8. It significantly reduces memory usage and accelerates inference, enabling the deployment of models on resource-constrained hardware [21–23]. However, a critical drawback of quantization is its inherent potential to degrade model accuracy, particularly when applied to complex architectures or in scenarios demanding high precision, such as segmentation [24–26]. Marginal accuracy degradation in brain tumor segmentation can lead to severe clinical implications, including misdiagnosis or suboptimal treatment planning. Moreover, complex architectures, including those improved with attention mechanisms, are often highly sensitive to this precision loss, as the effectiveness relies on subtle and precise feature weighting that is severely impacted by quantization-induced errors [27–29].

The existing literature often explores attention mechanisms for accuracy improvement and quantization for efficiency optimization largely in isolation or addresses the combination without fully mitigating the inherent compromises for highly sensitive medical applications. Few studies [30, 31] have systematically investigated how different attention mechanisms interact with various quantization strategies to achieve an optimal and practical balance, specifically for brain tumor segmentation, where both diagnostic precision as well as deployability are non-negotiable requirements. This gap shows a significant unresolved challenge in developing robust, accurate, and truly deployable deep learning models for clinical use. Therefore, a comprehensive understanding of this complex trade-off is crucial to

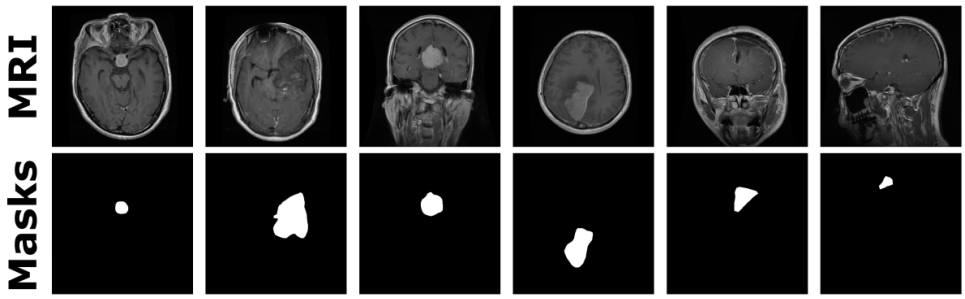

Fig. 1. Binary masks complementing Magnetic Resonance Imaging (MRI) tumor segmentation sets.

bridge the gap between theoretical advancements and practical clinical implementation.

B. Tumor Localization and Segmentation Validation in Brain MRI

In the research, further analysis aims to ensure that tumor predictions follow medical theories concerning the anatomical locations of specific tumor types, besides evaluating the segmentation results quantitatively using metrics such as IoU. For instance, gliomas are typically intra-axial, located in brain tissue, and frequently found in the frontal, temporal, or parietal lobes [32]. These tumors often have poorly defined margins due to the invasive nature and show a bright ring-like pattern post-contrast administration as an indicator of central necrosis [32]. Additionally, surrounding brain tissue swelling (peritumoral edema) is commonly observed on T2-weighted MRI as darker areas [33]. Meningiomas originate from the meninges and are extra-axial, often situated near the brain surface, such as along the falx cerebri or in the cranial fossae [34]. These tumors are characterized by well-defined, rounded, or oval shapes. They show homogeneous improvement after contrast administration. A distinguishing feature is the dural tail sign, which appears as a thickened edge adjacent to the tumor on MRI [35]. Following the discussion, peritumoral edema is usually less prominent compared to gliomas [33]. Pituitary tumors are located at the base of the brain in the sella turcica, just beneath the hypothalamus [36]. These tumors typically show uniform brightness post-contrast and may alter the sella turcica structure or exert pressure on the optic chiasm, potentially affecting vision [33].

II. Research Method

A. Brain Tumor Dataset for Classification and Segmentation

Medical image segmentation and classification heavily rely on high-quality datasets to develop robust and accurate deep learning models. Two widely used datasets in brain tumor studies are the brain tumor segmentation [37] and the brain tumor MRI dataset [38], both available on Kaggle. These datasets have been instrumental in advancing segmentation and classification methodologies, providing studies with comprehensive resources for evaluation as well as innovation.

The brain tumor segmentation dataset [37] provides high-resolution MRI images with detailed annotations, focusing on tumor boundary detection and segmentation (see Fig. 1). It is a benchmark dataset that is frequently referenced for developing and testing advanced segmentation architectures [39–42]. Moreover, the comprehensive nature of the dataset enables the exploration of complex tumor segmentation challenges, significantly contributing to improved segmentation accuracy in medical imaging.

Moreover, the brain tumor MRI dataset [38] is a key resource for classification tasks, categorizing images into glioma, meningioma, pituitary tumors, and no tumor (see Fig. 2). Its structured method supports the development of deep learning models for tumor classification, with numerous studies using the model to evaluate classification algorithms and propose novel architectures [43–46]. In addition, the dataset provides sufficient variability in tumor types and imaging conditions, making it suitable for evaluating model generalizability across diverse clinical scenarios. The availability of clearly defined classes facilitates comparative analysis between different deep learning architectures, driving continuous methodological improvements. Furthermore, the dataset has become a de facto benchmark for brain tumor classification research, ensuring reproducibility and enabling fair performance comparisons across studies. These characteristics establish the brain tumor MRI dataset as an indispensable resource for advancing computer-aided diagnosis in neuro-oncology.

In the research, the brain tumor segmentation dataset is the primary resource for training and evaluating the segmentation model. Moreover, the brain tumor MRI dataset is further used to assess the segmentation ability

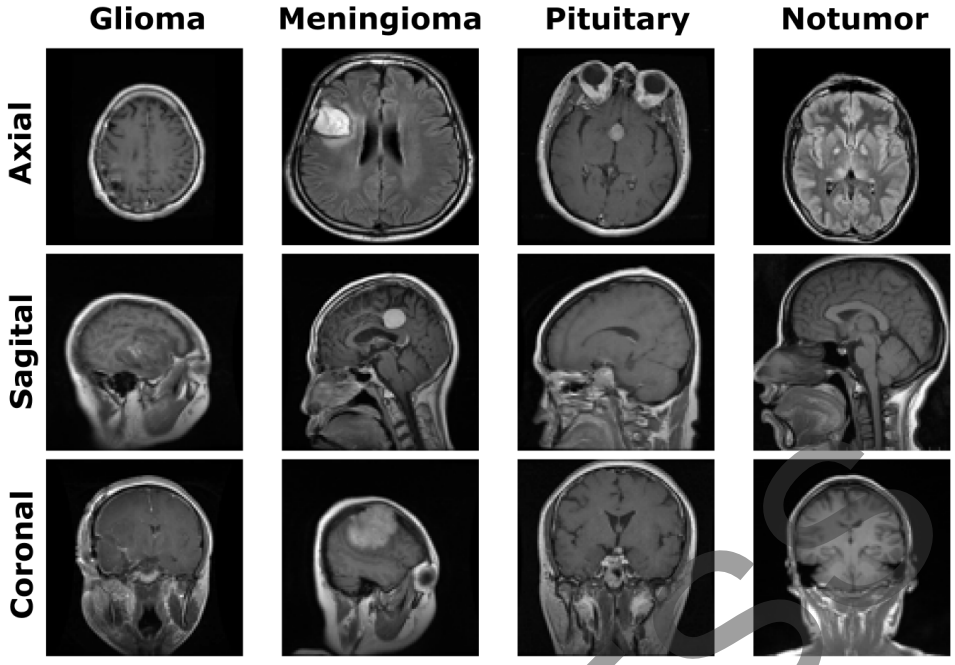

Fig. 2. Plane-wise Magnetic Resonance Imaging (MRI) images classified into four tumor types to identify the patient’s tumor type.

of the model to capture the distinct characteristics of glioma, meningioma, and pituitary tumors for validation. During the process, gliomas are evaluated for irregular shapes and peritumoral edema, meningiomas for rounded shapes and clear boundaries, and pituitary tumors for homogeneous shapes within the sella turcica. To enrich the dataset, “no tumor” images from the brain tumor MRI dataset are incorporated with empty segmentation masks. This augmentation ensures a balanced dataset, enabling the model to distinguish between healthy and tumorous brains during training effectively. This combined dataset method ensures a comprehensive evaluation of the performance of the model, showing the association with clinical expectations and the robustness in segmenting various tumor types.

B. U-Net and Its Development for Medical Segmentation

U-Net is a deep learning architecture widely used in medical image segmentation due to its ability to preserve critical spatial details through encoder-decoder structure and skip connections [1, 8, 47, 48]. The U-Net architecture, illustrated in the Fig. 3, represents one of the most influential fully convolutional networks in biomedical image segmentation. Its design follows a symmetric encoder–decoder structure, where the encoder progressively reduces spatial resolution through convolution and max-pooling operations, while simultaneously increasing the depth of feature representations. Each convolutional block consists of convolutional layers with batch normalization and Rectified Linear Unit (ReLU) activation, enabling efficient feature extraction and non-linear transformation of the input data.

The decoder mirrors this process by gradually reconstructing the segmentation map through a sequence of up-convolutions that restore spatial resolution. At each stage, the decoder integrates information from the encoder through skip connections, which concatenate high-resolution features from the contracting path with the upsampled features. These skip connections are critical for preserving fine-grained spatial information that might otherwise be lost during downsampling, thereby ensuring accurate boundary reconstruction of the segmented regions. The final output is generated by a 1×1 convolution, which maps the feature representation to the desired segmentation classes. Despite its success, U-Net faces limitations in capturing complex features in challenging cases, such as irregular brain tumors. To address these challenges, Res-UNet combines spatial feature learning of U-Net with residual blocks of ResNet. It improves training stability and enables deeper learning without gradient degradation [3, 14].

Res-UNet has shown superior performance in medical segmentation tasks, including brain tumor seg-

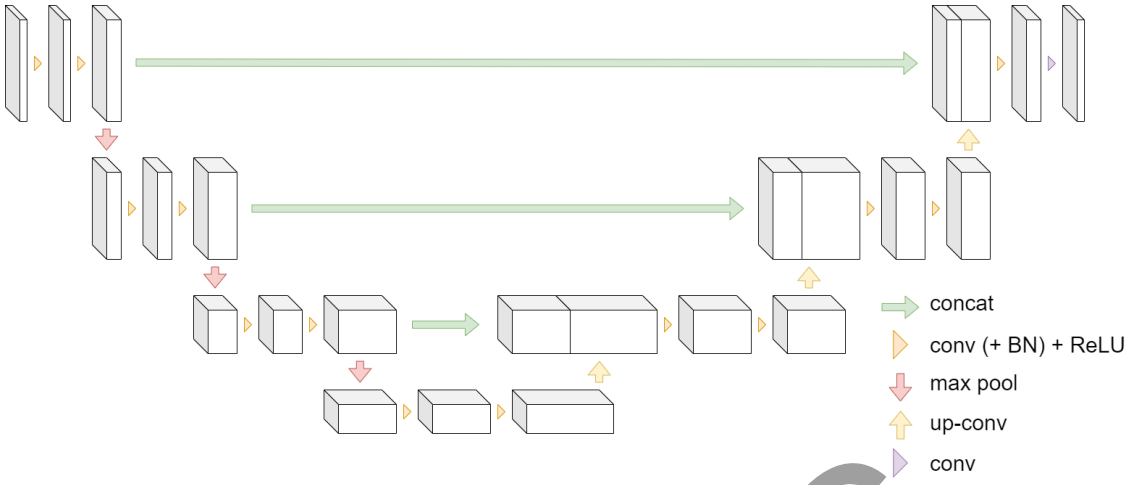

Fig. 3. Example of U-Net architecture. Note: Concatenate (concat), BatchNormalization (BN), Rectified Linear Unit (ReLU), and Convolution (conv.).

mentation, by effectively capturing contextual information and handling ambiguous tumor boundaries [49]. Advanced implementations, such as nnU-Net, further optimize Res-UNet for medical data without extensive manual adjustments [2]. The research uses Res-UNet as the foundation to achieve more accurate and robust brain tumor segmentation, surpassing the capabilities of the standard U-Net.

C. Attention Mechanisms in Deep Learning for Segmentation

Attention mechanisms have appeared as major innovations in deep learning for tasks requiring precision, including brain tumor segmentation. By enabling models to focus on critical regions while suppressing irrelevant information, attention mechanisms improve segmentation accuracy, particularly for challenging features such as tumor boundaries [14, 50]. In the research, the Res-UNet architecture is enhanced with three attention mechanisms: AG, SE, and CBAM. These mechanisms improve the ability of the model to prioritize key features, each contributing uniquely to segmentation performance.

The AG-augmented U-Net architecture improves upon the standard U-Net by embedding AG modules within the skip connections, designed to direct the focus of the model on essential regions, such as tumors, by filtering out irrelevant information (Fig. 4). These modules operate as feature selectors, ensuring that only task-relevant encoder representations are propagated to the decoder, thereby enhancing focus on diagnostically salient regions such as tumor boundaries in MRI. Each AG is driven by two inputs: a gating signal g_i, derived from the decoder to provide contextual guidance, and the encoder feature map x. Both inputs undergo linear transformations via convolutional operations, yielding Eq. (1). The W_φ and W_0 denote learnable kernels. The θ_x is the transformed encoder feature map after convolution, containing spatially reduced but semantically rich information. Then, φ_g is the transformed gating signal, which carries the semantic context from the decoder in a lower-dimensional form.

$$\begin{aligned} \varphi_g &= W_\varphi * g, \\ \theta_x &= W_0 * x. \end{aligned} \tag{1}$$

The intermediate representation is then computed as Eq. (2). The f denotes the element-wise addition of both transformed features, combining spatial and contextual information before attention weighting. The combined features f are processed through a Leaky ReLU activation, introducing non-linearity while retaining negative responses with a small slope (0.1 f for $f \leq 0$). Subsequently, the features are refined via an additional convolution and passed through a sigmoid activation to generate the attention coefficients, as seen in Eq. (3). The α (or f after the sigmoid) represents the attention coefficients indicating the relative importance of each spatial location in the encoder feature map. Then, W_f is the convolution kernel used to process the activated signal f and generate the raw attention coefficients.

$$f = \theta_x + \varphi_g, \tag{2}$$

$$\alpha = \sigma(W_f * f),\ \alpha\, \epsilon\, (0, 1). \tag{3}$$

These coefficients function as soft masks that scale the encoder activations through element-wise multiplication in Eq. (4). Through this mechanism, AG modules are particularly useful for complex image

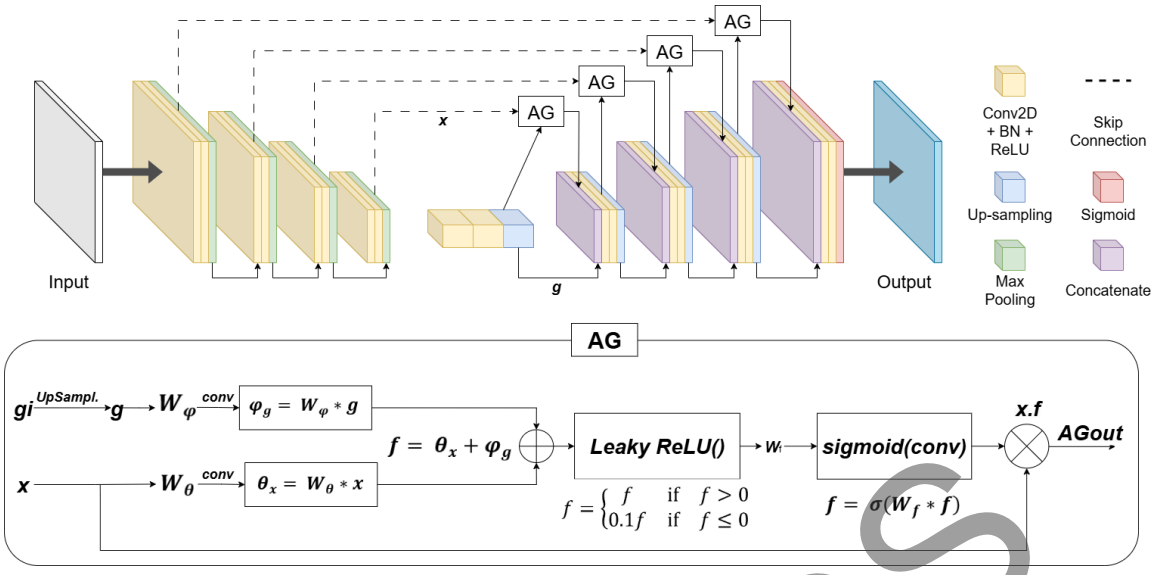

Fig. 4. Attention Gate (AG) implementation in U-Net architecture. Note: BatchNormalization (BN), Rectified Linear Unit (ReLU), and Convolution (conv.).

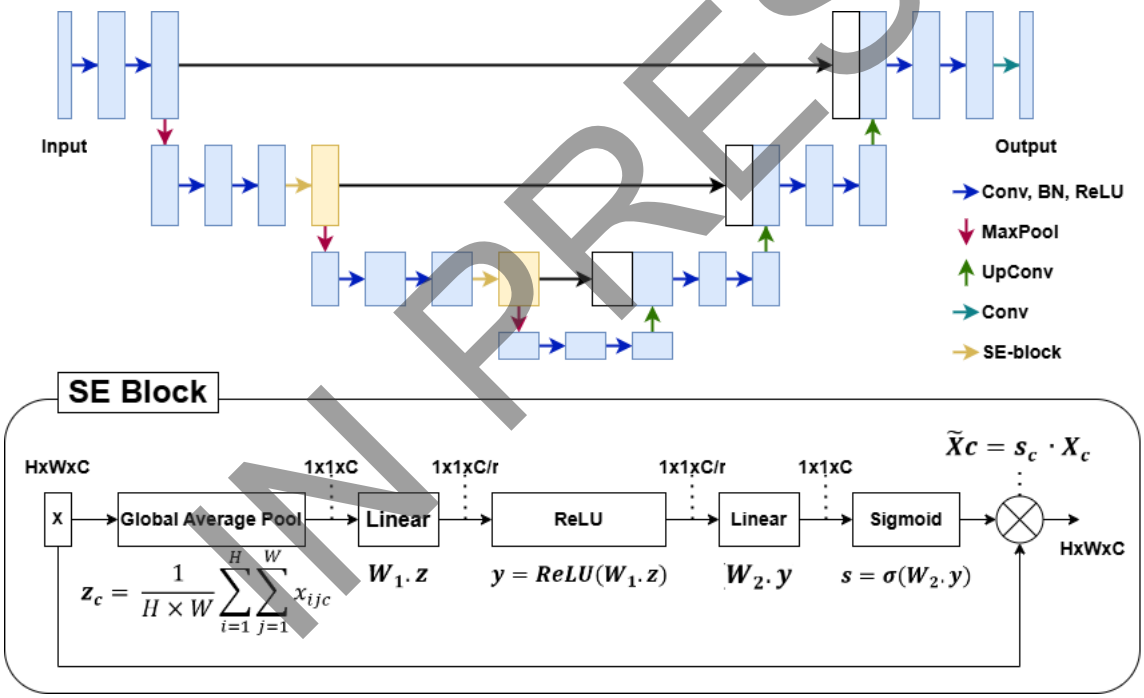

Fig. 5. Squeeze and Excitation (SE) Block implementation in U-Net architecture. Note: BatchNormalization (BN), Rectified Linear Unit (ReLU), and Convolution (conv.).

segmentation tasks, such as brain tumor MRI, where object boundaries are often blurred or the tumor is small. The AG dynamically calculates attention coefficients based on encoder and decoder features, determining which areas should be stressed or suppressed. It enables more accurate identification of critical regions, such as tumor boundaries, in MRI images [4, 14]. The filtering mechanism of AG allows models to focus dynamically on high-probability target regions, improving segmentation accuracy for ambiguous areas.

$$AG_{out} = x \,.\, \alpha. \tag{4}$$

The architecture (Fig. 5) illustrated integrates the SE Block into the U-Net framework, enabling channel-

wise feature recalibration during the encoding and decoding processes [5, 15]. Within the U-Net backbone, the conventional operations–convolution with batch normalization and ReLU activation (Conv., BN, ReLU), max pooling for downsampling, and transposed convolution (UpConv) for upsampling–are preserved. The SE Block modules are embedded after selected convolutional layers, augmenting the representational power by explicitly modeling interdependencies among feature channels.

The SE Block functions in two key stages: squeeze and excitation. In the squeeze stage, global spatial information is aggregated through Global Average Pooling across each channel, yielding a channel descriptor in Eq. (5). The x_{ijc} denotes the feature response at spatial location (i, j) in channel c, and H and W are the feature map dimensions. This descriptor z_c captures the global context of each channel. In the excitation stage, the aggregated vector z is passed through two fully connected (FC) layers with a bottleneck ratio r. The first layer performs a dimensionality reduction, followed by a ReLU non-linearity in Eq. (6). It has $z = [z_1, z_2, \ldots, z_C]$ as a C-dimensional vector, one value per channel, describing how much information each channel carries globally.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{ijc}, \quad (5)$$

$$y = \text{ReLU}(W_1 \,.\, z). \quad (6)$$

The second FC layer restores the dimensionality, followed by a sigmoid activation to produce channel-wise weights in Eq. (7). The s is sigmoid for activation function, mapping values to the range [0, 1], where each value s_c indicates the relative importance of channel c, and y is previous ReLU described in Eq. (6). The $s\epsilon(0,1)^C$ acts as attention coefficients. These coefficients are then applied to the original feature maps via element-wise multiplication in Eq. (8) where each channel X_c is adaptively reweighed according to its importance. Then, $\widetilde{X_c}$ represents the recalibrated output feature map, obtained by multiplying each channel X_c of the input feature map by its corresponding attention weight s_c.

$$s = \sigma(W_2 \,.\, y), \quad (7)$$

$$\widetilde{X_c} = s_c \,.\, X_c. \quad (8)$$

This mechanism allows the network to emphasize more informative channels while suppressing less relevant ones, thereby refining feature representation with minimal computational overhead. By integrating SE Blocks into U-Net, the model enhances its sensitivity to subtle but diagnostically significant features in medical images. It leads to improved segmentation accuracy while maintaining efficiency.

The CBAM enhances feature representation by sequentially applying two complementary attention mechanisms: channel attention and spatial attention (Fig. 6). These modules are lightweight and can be seamlessly integrated into convolutional neural networks, including U-Net variants, without significant computational overhead. By refining features along both channel and spatial dimensions, CBAM allows the network to learn what and where to emphasize, thereby improving interpretability and performance in tasks such as medical image segmentation [6, 15, 16]. The Channel Attention Module (CAM) focuses on identifying the relative importance of each feature channel. Given an intermediate feature map $F\epsilon\mathbb{R}^{H\times W\times C}$, CBAM computes both average-pooling and max-pooling across spatial dimensions, producing two channel descriptors in Eq. (9). The F is the input feature map with dimensions $H \times W \times C$, where H and W are spatial dimensions, and C is the number of channels. Global Average Pooling and Global Max Pooling are applied independently across the spatial dimensions of F.

$$F^c_{avg} = AvgPool(F),\ F^c_{max} = MaxPool(F). \quad (9)$$

These descriptors are forwarded through a shared Multi-Layer Perceptron (MLP) with one hidden layer to capture non-linear channel dependencies. The two outputs are then combined element-wise, as seen in Eq. (10). The σ is the sigmoid activation. The resulting channel attention map $M_c(F)$ is multiplied with the original feature map to yield channel-refined features in Eq. (11).

$$M_c(F) = \sigma(MLP(F^c_{avg}) + MLP(F^c_{max})), \quad (10)$$

$$F' = M_c(F) \bigotimes F. \quad (11)$$

The Spatial Attention Module (SAM) further emphasizes where informative features are located. Using the channel-refined feature map F', average-pooling and max-pooling are applied along the channel axis, generating two 2D maps that highlight complementary spatial cues. These maps are concatenated and convolved with a 7×7 kernel, as seen in Eq. (12). The $f^{7\times7}$ denotes convolution. The resulting spatial attention map $M_s(F')$ is element-wise multiplied with F' to generate the final refined output (see Eq. (13)).

$$M_s(F') = \sigma(f^{7\times7}([AvgPool(F'); MaxPool(F')])), \quad (12)$$

$$F'' = M_s(F') \bigotimes F'. \quad (13)$$

By sequentially combining CAM and SAM, CBAM

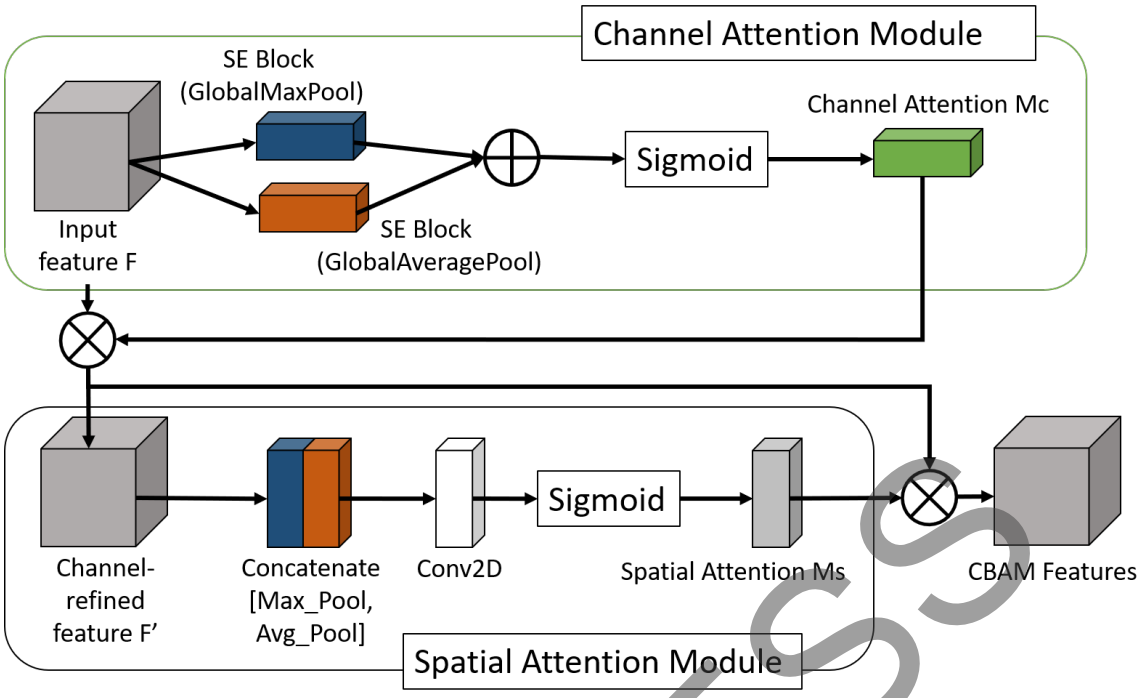

Fig. 6. Convolutional Block Attention Module (CBAM) implementation in U-Net architecture. Note: Squeeze-and-Excitation (SE) Block, Channel Attention Module (Mc), and Channel Attention Spatial (Ms).

adaptively recalibrates features both across channels and within spatial regions. This dual refinement enables the model to focus on diagnostically relevant patterns while suppressing background noise. In medical image segmentation tasks, such as brain tumor analysis, this approach provides a more comprehensive feature enhancement compared to channel-only or spatial-only mechanisms, making it highly effective for capturing subtle lesion characteristics.

D. Dynamic Range Quantization

The DRQ is an optimization method that reduces the numerical precision of deep learning model parameters, mapping full-precision floating-point (FP32) values into lower-precision integer (INT8) representations while preserving the relative dynamic range. This method significantly reduces model size and improves inference efficiency without substantially degrading performance. DRQ is particularly beneficial for resource-constrained medical environments, where deep learning models need to run efficiently on edge devices.

The conversion from FP32 to INT8 is performed by scaling and rounding the weight values according to the dynamic range of the original data. The quantized weight w_q is obtained using Eq. (14). Then, rearranging the equation to solve for weight w is shown in Eq. (15). The w_{min} and w_{max} are the minimum and maximum values of the original FP32 weights. Then, $w_{qmin} = -127$ and $w_{qmax} = 127$ represent the fixed dynamic range for INT8 quantization.

$$\frac{w - w_{min}}{w_{max} - w_{min}} = \frac{w_q - w_{qmin}}{w_{qmax} - w_{qmin}}, \tag{14}$$

$$w_q = round(\frac{(w - w_{min})(w_{qmax} - w_{qmin})}{w_{max} - w_{min}} + w_{qmin}). \tag{15}$$

Unlike static quantization, where scale factors are predetermined, DRQ dynamically determines scale factors per layer based on observed activation ranges, ensuring better adaptability to various data distributions. Quantization can achieve significant memory reduction (e.g., 8-fold) with minimal performance loss, often under 2% in metrics such as dice score [27]. It shows the practicality of quantization for deploying deep learning models in resource-constrained medical environments.

E. U-Net and Its Development for Medical Segmentation

The performance evaluation of models in medical image segmentation, such as brain tumor segmentation, typically includes several metrics to assess the

accuracy and efficiency of predictions. In the context of deep learning, these metrics measure how well the model predicts tumor regions compared to ground truth labels [51, 52]. The research uses IoU as the primary metric to evaluate segmentation performance. IoU, also known as the Jaccard Index, quantifies the similarity between the predicted area and the ground truth by calculating the ratio of the intersection to the union. IoU is widely used in segmentation tasks because it provides a conservative assessment compared to other metrics, such as the Dice coefficient, and shows the sensitivity of the model to small overlaps between predictions and ground truth [53]. Equation (16) shows the formula of IoU. The A represents the set of pixels (or voxels, in the case of 3D MRI) predicted by the model as belonging to the tumor region, and B represents the set of pixels corresponding to the ground truth tumor annotation provided by expert radiologists. The numerator $|A \cap B|$ measures the correctly predicted tumor pixels (true positives), while the denominator $|A \cup B|$ accounts for all pixels labeled as tumor in either prediction or ground truth, thereby penalizing both false positives and false negatives.

$$IoU = \frac{|A \cap B|}{|A \cup B|}. \tag{16}$$

F. Proposed Method

The research introduces a novel framework for brain tumor segmentation, combining Res-UNet with advanced attention mechanisms and DRQ to achieve both high accuracy and computational efficiency. The innovation is in the dual focus of improving segmentation precision through attention mechanisms, namely AG, SE Block, and CBAM, while ensuring practical deployability in resource-constrained environments through quantization. The workflow of the proposed methodology is shown in Fig. 7.

The research leverages two distinct datasets, both sourced from Kaggle and containing MRI-based brain images. The classification dataset is designed for categorizing various brain conditions and comprises a total of 7,023 images. These are distributed among 1,621 glioma, 1,645 meningioma, 2,000 no-tumor, and 1,757 pituitary tumor samples. Consequently, the segmentation dataset focuses specifically on marked-out tumor regions. It features 3064 tumor images with associated segmentation masks, alongside 2,000 no-tumor images with empty masks. Moreover, the masks are crucial for training the model to differentiate healthy from tumorous areas effectively.

Before model training, all images from both datasets have passed through essential preprocessing steps to standardize the size and format. It includes resizing all input images and the corresponding masks to a uniform dimension of 224×224 pixels, as the standard input size for the models. During the process, pixel intensity values for both images and masks are normalized, divided by 255, and scaled to a range of 0.0 to 1.0 to ensure consistent input for the neural network. Masks are specifically loaded in grayscale mode to represent the binary segmentation targets. Concerning "no tumor" samples in the segmentation dataset, empty masks are specifically generated to provide corresponding target labels during training. For robust evaluation, the combined dataset is then rigorously divided into training (60%), validation (20%), and testing (20%) subsets. Moreover, data batches for training and validation are efficiently handled using a custom image_mask_generator function, which loads images as well as the masks from specified directories in PNG format and produces batches of normalized image-mask pairs.

The foundation of the methodology is built upon Res-UNet architecture, which seamlessly incorporates the spatial feature extraction capabilities of U-Net with the powerful residual learning of ResNet. The base Res-UNet model, serving as the foundational architecture, adopts a symmetric encoder-decoder structure. Its encoder path is derived from a pre-trained ResNet50V2 backbone (fine-tuned on ImageNet), which extracts hierarchical features. Moreover, Key feature maps from intermediate layers of the backbone (conv1_conv, conv2_block3_1_conv, conv3_block4_1_conv, conv4_block5_out, and post_bn for the deepest features) are used as skip connections. The decoder path progressively upsamples these features to reconstruct the segmentation mask, mirroring the downsampling stages of the encoder. Major building blocks include a conv_block (a 3×3 Conv2D layer with BatchNormalization (BN) and Leaky Rectified Linear Unit (LeakyReLU) activation) and a res_block (applying two conv_blocks with a residual shortcut connection, optionally followed by BN and LeakyReLU. Regarding the process, the decoder reconstructs the segmentation map through a series of Conv2DTranspose layers for upsampling, followed by concatenation with the corresponding encoder skip connections, and subsequent processing by the res_block. During the process, dropout layers are strategically placed to mitigate overfitting. The final output layer is a Conv2D with a 1×1 kernel and sigmoid activation, producing a single-channel probability map. The complete architecture is shown in Fig. A1 in Appendix.

The three distinct attention mechanisms are sys-

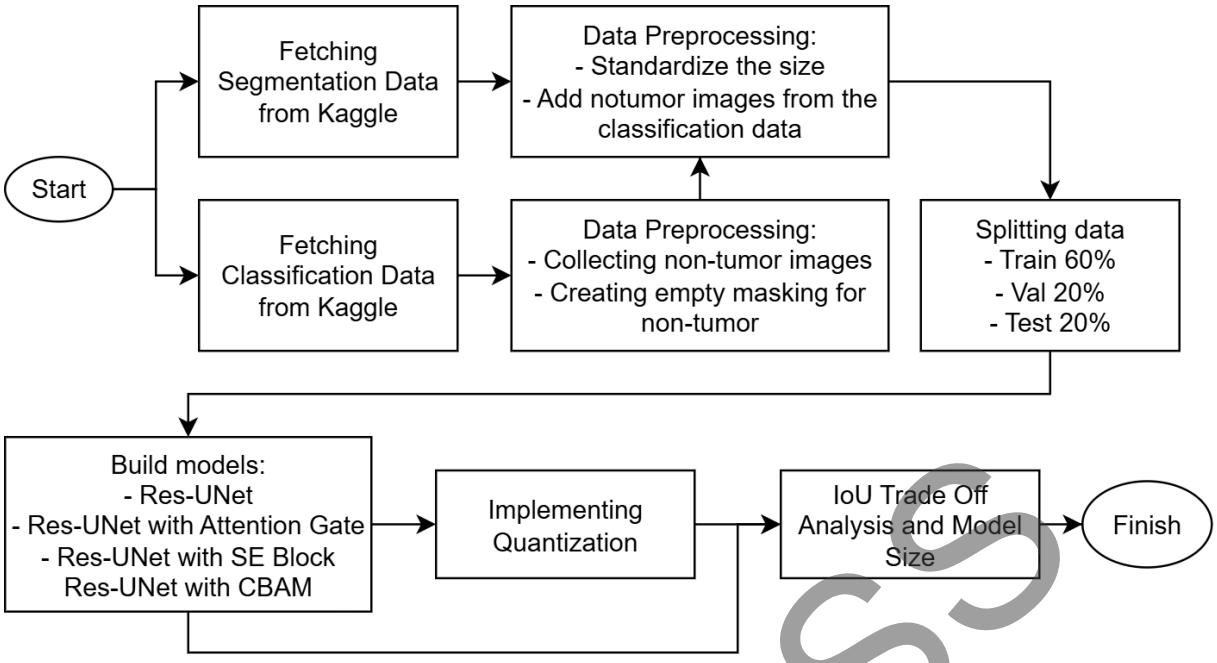

Fig. 7. Proposed methodology. Note: Squeeze-and-Excitation (SE) Block, Convolutional Block Attention Module (CBAM), Val: Validation, and Intersection over Union (IoU).

tematically incorporated to improve segmentation performance. First, AG is incorporated into each skip connection between the encoder and decoder paths. The attention_gate module receives a feature map from the encoder (x) and a gating signal from the decoder (g). It applies 1×1 convolutional layers to both inputs to associate the channels, sums, passed through 'LeakyReLU' activation, and another 1×1 Conv2D layer with sigmoid activation in generating a spatial attention map. Subsequently, the map is element-wise multiplied with the original encoder feature map. This process adaptively filters out irrelevant features from the low-level feature maps of the encoder, ensuring that only diagnostically relevant features contribute optimally to the decoder path. AG is selected because it is computationally lightweight and highly effective in showing diagnostically relevant regions, making the process suitable for resource-constrained deployment scenarios [54, 55].

Second, SE Block is incorporated at two major locations, on the skip connections (layer5, layer4, etc., from the backbone) and in the res_blocks. The se_block operated by first using GlobalAveragePooling2D to compress spatial information for each channel (squeeze operation). It is followed by an excitation operation using two Dense layers (with ReLU and Sigmoid activations, and a reduction ratio of 16) to generate channel-wise attention weights. Subsequently, these weights are element-wise multiplied with the input feature map, allowing the model to learn the interdependence between channels and accentuate more informative ones. SE Block is selected because the features provide an efficient channel refinement mechanism with minimal parameter overhead, offering a good trade-off between improved feature representation and computational cost [56].

Third, CBAM is incorporated similarly to SE Block, both in the skip connections and in the residual blocks. The cbam_block sequentially applies attention across two dimensions, namely channel and spatial. During the process, the CAM uses both GlobalAveragePooling2D and GlobalMaxPooling2D to aggregate spatial information, which is then passed through a shared MLP (with a reduction ratio of 8) to generate channel attention weights. The SAM takes the output of the CAM, performs average-pooling and max-pooling along the channel axis, and concatenates and convolves the features with a single 7×7 Conv2D layer to produce a spatial attention map. Both attention maps are later applied sequentially to refine the feature maps, enabling the model to learn "what" is important (channel-wise) and "where" it is important (spatial-wise) in the feature maps. Additionally, CBAM is included because it combines both spatial and channel attention, providing a more comprehensive feature improvement. Despite being slightly heavier computa-

tionally, it makes the model valuable for systematic comparison with AG and SE. The impact of these mechanisms on segmentation accuracy is rigorously evaluated using IoU [57].

The novelty is extended further with the application of DRQ, which optimizes the trained models for deployment by reducing the precision of weights and activations. DRQ is a Post-Training Quantization (PTQ) method applied to the trained 32-bit floating-point (FP32) models. Its primary objective is to convert all quantizable model parameters (weights) and intermediate activations to lower-precision formats, specifically 8-bit integers (INT8), significantly reducing memory footprint (typically achieving up to 75% reduction) and accelerating computational demand. The process includes observing the dynamic range (minimum and maximum values) of tensors during a small calibration step. Following the discussion, a representative subset of the input data (e.g., 100–500 images from the validation set) is fed through the trained FP32 model to collect these range statistics. Based on these observed min/max values, scaling factors and zero-points are computed for each tensor, which are then used to map the original FP32 values to the fixed-point INT8 range. This method allows the quantized model to operate with reduced precision without requiring specific hardware accelerators, making the model highly suitable for deployment on resource-constrained hardware, such as mobile devices, embedded systems, or standard hospital workstations. DRQ is selected for its simplicity and the ability to apply optimization without requiring model retraining, which is crucial for rapid deployment scenarios. In contrast, Quantization-Aware Training (QAT) potentially produces higher accuracy by incorporating quantization effects during the training phase. The quantization process is typically implemented using functionalities provided by standard machine learning frameworks such as TensorFlow Lite converter.

Concerning the training of all Res-UNet variants (base, AG, SE, CBAM), the researchers compile models using the Adam optimizer. A learning rate of 0.001 is initially set, and training is performed for 50 epochs with a batch size of 16. During the process, the binary cross-entropy loss function is used, given the binary nature of the segmentation task (tumor vs. non-tumor). Model performance is monitored using the 'val_binary_io_u' metric on the validation set, with mode='max' to track improvements. Training incorporates two major callbacks to ensure optimal and stable learning as follows:

1) EarlyStopping is configured with a patience of 7 epochs, monitoring 'val_binary_io_u' in mode='max', and set to restore the weights of the best-performing model,
2) ReduceLROnPlateau is used to dynamically adjust the learning rate, reducing it by a factor of 0.2 after 4 epochs without improvement, with a minimum learning rate of 1e-6 to prevent excessive decay.

All models are trained and evaluated on a single NVIDIA Tesla M10 GPU to maintain a consistent computational environment. Inference latency is measured as the average per image over 500 forward passes using the same GPU to ensure reproducibility and to avoid warm-up effects. This setup ensures that the reported latency comparisons among Res-UNet variants (base, AG, SE, CBAM) and the quantized versions are hardware-consistent and directly comparable.

The methodology concludes with an extensive evaluation of the trade-offs between segmentation accuracy and computational efficiency. The experimental results demonstrate that the proposed approach is able to identify an optimal balance between these two objectives, ensuring that the model remains both accurate and efficient. This balance is particularly important in the context of medical image segmentation, where diagnostic precision cannot be compromised while computational resources are often limited. By integrating state-of-the-art attention mechanisms with quantization strategies, the proposed framework establishes a strong benchmark for practical deployment. These findings can highlight the potential of the method to advance the field of medical imaging and support real-world clinical applications under resource-constrained environments.

III. Results and Discussion

The research investigates brain tumor segmentation using Res-UNet variants improved with AG, SE Block, and CBAM. The evaluation comprises training loss, validation loss, IoU, and test performance metrics, such as Mean IoU, model size, and inference time. These metrics provide a comprehensive assessment of both predictive accuracy and computational efficiency, enabling a fair comparison between different model variants. In particular, the inclusion of model size and inference time is critical to demonstrate the practicality of the approach for real-world clinical environments, where hardware resources may be limited. Furthermore, by monitoring training and validation loss trends, the study ensures that the models achieve generalization without overfitting, thus strengthening the reliability of the results.

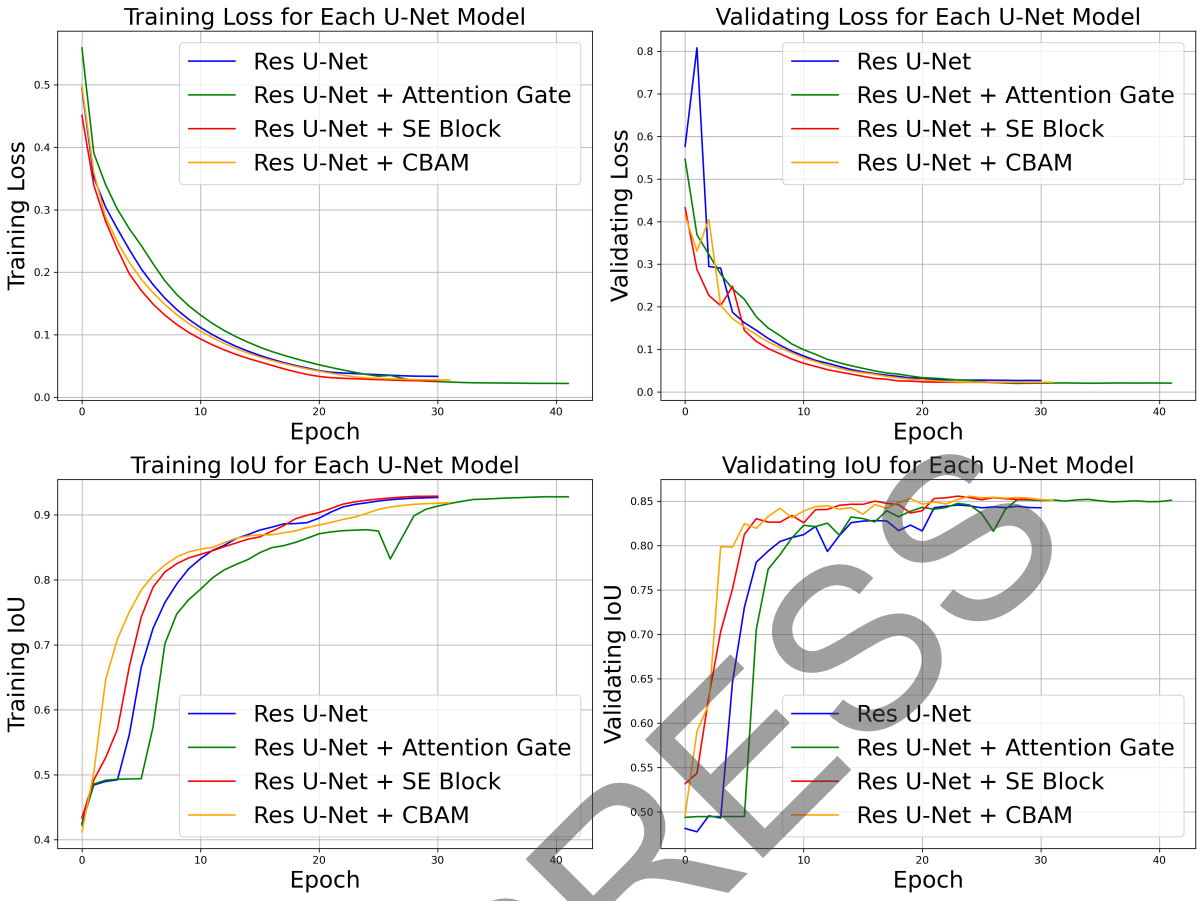

Fig. 8. Comparison of training and validation across U-Net variants. Note: Intersection over Union (IoU), Squeeze-and-Excitation (SE) Block and Convolutional Block Attention Module (CBAM).

A. *Evaluation of Training and Validation*

The training and validation curves provide valuable insights into how each attention mechanism influences model convergence, generalization, and stability during learning. As shown in Fig. 8, Res-UNet + SE Block achieves the lowest training loss, followed closely by CBAM, while the baseline Res-UNet converges to a higher loss. This trend reflects that both SE Block and CBAM facilitate more efficient feature learning by adaptively reweighting informative channels and spatial features in the case of CBAM, thereby allowing the model to focus on diagnostically relevant regions while suppressing redundant activations. The SE Block, in particular, performs a channel-wise recalibration that enhances inter-channel dependencies, leading to improved representational capacity and faster convergence. The result is consistent with previous findings [5, 56] that channel attention mechanisms accelerate training by stabilizing gradient propagation across layers.

The CBAM variant shows comparably low training loss but slightly slower convergence than SE Block. This behavior can be attributed to CBAM’s two-stage attention refinement, which involves both channel and spatial attention, introducing additional computations and dependencies that delay convergence in early epochs but yield stronger spatial awareness later in training [6, 57]. The baseline Res-UNet, lacking any attention mechanism, relies purely on residual learning, which still improves gradient flow but cannot selectively suppress irrelevant background noise. As a result, it learns more slowly and retains a higher steady-state loss.

The AG variant demonstrates an interesting trade-off. Its training loss decreases more gradually, and its final IoU stabilizes around 0.90, which is slightly below that of SE Block and CBAM. This slower convergence is expected because AG selectively filters encoder features based on decoder context–effectively gating gradient flow to emphasize semantically relevant regions only [14, 54]. While this controlled feature selection may reduce the volume of information

TABLE I
COMPARISON OF RES-UNET AND THE VARIANTS IN TERMS OF FINAL MEAN INTERSECTION OVER UNION (IOU), MODEL SIZE, AND INFERENCE TIME.

Model	Final Mean IoU	Model Size (MB)	Interface Time (s)
Res-UNet	0.8322	126.4877	0.3143
Res-UNet (DRQ)	0.8317	32.1887	0.1973
Res-UNet + AG	0.8455	127.2101	0.3131
Res-UNet + AG (DRQ)	0.8451	32.3863	0.2055
Res-UNet + SE	0.8406	127.0890	0.3171
Res-UNet + SE (DRQ)	0.8384	32.3802	0.1996
Res-UNet + CBAM	0.8359	128.8157	0.3245
Res-UNet + CBAM (DRQ)	0.8286	32.8767	0.2045

Note: Squeeze-and-Excitation (SE) Block, Convolutional Block Attention Module (CBAM), Attention Gate (AG), and Dynamic Range Quantization (DRQ).

passed through the network, it also mitigates overfitting by preventing the model from memorizing irrelevant spatial patterns. Consequently, AG sacrifices a small amount of training speed for more robust generalization, which is confirmed by its lowest validation loss among all variants.

Validation performance further clarifies these dynamics. Res-UNet + AG maintains the smallest validation loss and exhibits the most stable curve, signifying excellent generalization and minimal overfitting. This result aligns with prior evidence that attention gating improves localization accuracy in medical segmentation tasks by suppressing irrelevant activations [54, 55]. SE Block and CBAM both achieve strong validation IoU (≈ 0.85), but their slightly higher validation losses suggest mild over-adaptation to training data due to increased model complexity. The baseline Res-UNet, with the highest validation loss (≈ 0.84 IoU), confirms that residual connections alone are insufficient for precise tumor boundary discrimination, especially in cases with subtle contrast variations.

Overall, the comparative analysis indicates that integrating attention mechanisms not only enhances learning efficiency but also contributes to better bias-variance balance, a hallmark of models that generalize well to unseen data [52]. AG's adaptive feature gating yields the most stable validation performance, SE Block offers the fastest convergence, and CBAM provides a richer spatial context, albeit with a slightly higher computational cost. These complementary behaviors underscore that attention mechanisms meaningfully improve both the learning dynamics and the clinical reliability of Res-UNet in brain-tumor segmentation.

B. Testing Evaluation for Res-UNet Models

Table I compares the quantitative performance of all Res-UNet variants in terms of segmentation accuracy (Mean IoU), model size, and inference time. The results highlight a consistent trade-off between representational richness and computational efficiency across architectures. The DRQ demonstrates remarkable effectiveness by significantly reducing the memory footprint while preserving segmentation quality. The quantized models show up to 75% reduction in model size (Res-UNet + AG decreases from 127.21 MB to 32.39 MB), and inference speed improves by approximately 37% (0.314 s to 0.197 s per image). Importantly, the drop in mean IoU is minimal (less than 0.0004), confirming that post-training quantization introduces negligible accuracy degradation for segmentation tasks. This observation aligns with previous findings [27, 29] that DRQ can retain over 98% of baseline performance when the network maintains stable activation distributions and batch normalization layers are properly calibrated.

Among all models, Res-UNet + AG achieves the most favorable balance of accuracy, compactness, and inference speed (IoU = 0.845, 32.39 MB, 0.205 s). This advantage can be attributed to the computational design of AG modules. AG operates primarily through lightweight 1×1 convolutions and element-wise multiplications applied to skip connections, effectively filtering irrelevant encoder features before they are merged into the decoder path. This focused processing minimizes unnecessary computations while improving sensitivity to diagnostically relevant regions. The efficiency and interpretability of AGs have been previously emphasized in medical imaging research, such as [54], where attention gating is shown to reduce computational load and improve lesion localization by focusing only on salient features.

The SE Block variant also performs competitively, achieving a mean IoU of 0.841 and an inference time of 0.317 s. Its slightly higher computational demand stems from the use of global pooling and two fully connected layers in each block, which increases channel dependency modeling at the expense of latency [56]. Nevertheless, SE enhances channel sensitivity and gradient stability, contributing to smoother convergence and high-quality feature abstraction. This result aligns with previous research [56] that SE recalibration significantly boosts discriminative power, particularly in networks dealing with subtle intensity variations as found in MRI scans.

CBAM, which combines both channel and spatial attention, produces rich feature representations but exhibits the slowest inference (0.3245 s) and the largest model size (128.8 MB). This result is expected, as the CBAM processes both average- and max-pooled features, followed by a 7×7 convolution to generate spatial attention maps. While such dual refinement im-

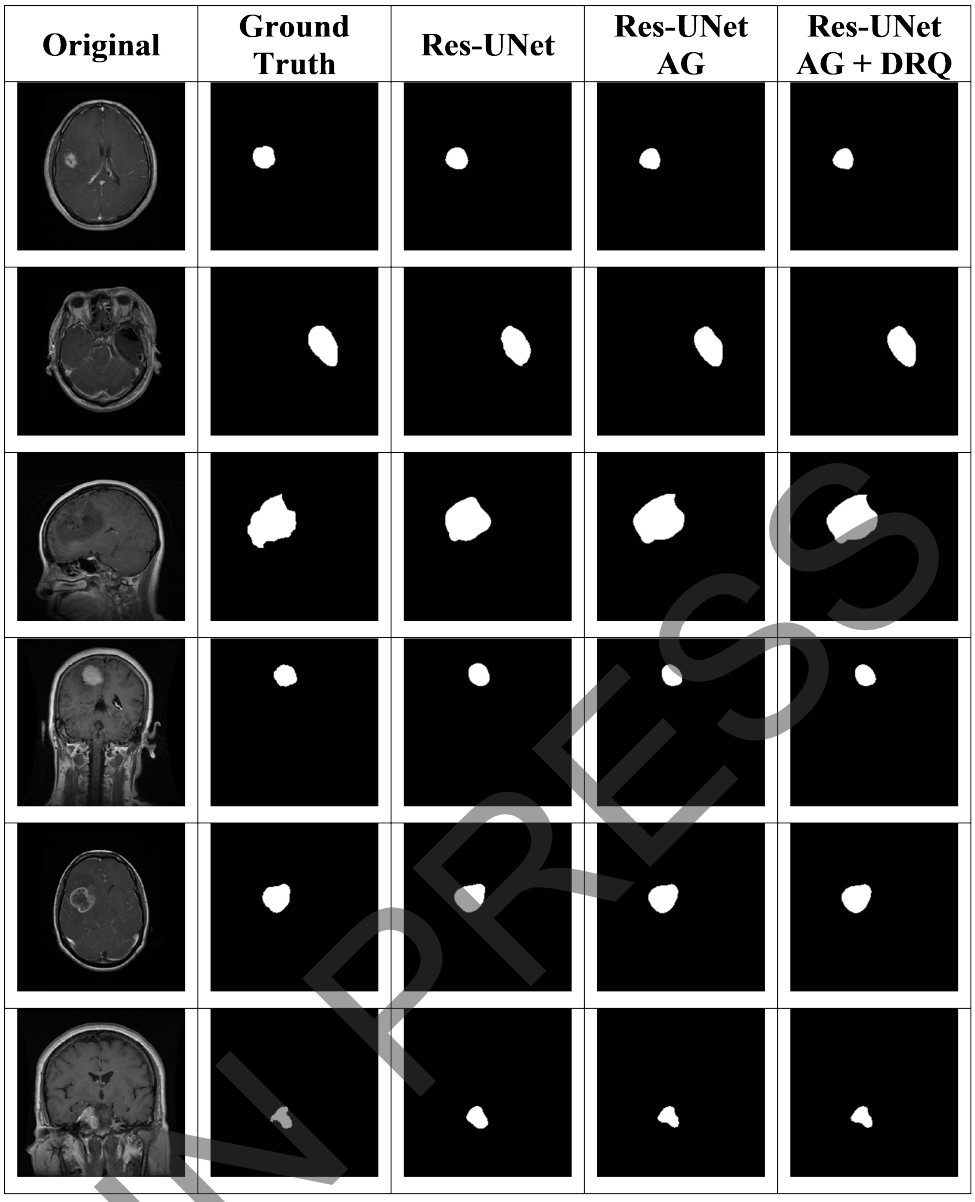

Fig. 9. Visualization of segmentation results for each model. Note: Attention Gate (AG) and Dynamic Range Quantization (DRQ).

proves lesion boundary detection and contextual understanding, it adds measurable computational overhead. Prior studies have also reported that CBAM enhances segmentation fidelity but introduces additional latency due to its two-stage attention process [16, 57]. Therefore, CBAM represents a trade-off favoring segmentation precision and interpretability over pure efficiency.

Quantization consistently preserves the accuracy ranking across all variants, suggesting that attention mechanisms increase the network's resilience to reduced numerical precision. The attention modules appear to regularize the feature distribution, reducing sensitivity to quantization noise, an effect similarly reported previously [24] that structured attention layers stabilize activation variance during low-bit quantization. Consequently, Res-UNet + AG (DRQ) remains the top-performing model even after quantization, validating its suitability for deployment of edge devices and standard hospital hardware where inference speed and memory efficiency are critical.

Representative segmentation outputs, showing the qualitative performance of the models, are shown in Fig. 9. These results signify the practical viability of attention-improved Res-UNet models with quantization for deployment in resource-constrained environments. The visualizations clearly demonstrate that

the model predictions align closely with the ground truth annotations, highlighting the ability of the models to capture tumor boundaries with high fidelity. Even in challenging cases with irregular tumor shapes or low-contrast regions, the models exhibit robustness by preserving critical structural details. Furthermore, the qualitative comparison underscores how attention-enhanced architectures reduce false positives and improve delineation in ambiguous regions, validating the consistency of the quantitative improvements observed in IoU.

C. Analysis of Segmentation Appropriateness

The research evaluates MRI segmentation results using the Res-UNet model with quantization. In addition to metrics such as IoU, the outcomes are evaluated for their association with medical theories regarding the anatomical positions and characteristics of tumor types. The segmentation appropriately captures expected patterns for glioma, meningioma, and pituitary tumors. For glioma, the model appropriately identifies ring improvement around the main lesion, a hallmark in MRI, and detects peritumoral edema, often appearing as dark areas due to the invasive nature of the tumor. Despite unclear boundaries caused by the spread of the glioma into brain tissues, the segmentation follows theoretical expectations, effectively capturing its complex invasive patterns and edema effects. Figure A2 in Appendix includes a sample showing the capability of the models during the process of the analysis.

In glioma segmentation, the model successfully identifies the characteristic ring enhancement pattern surrounding the necrotic tumor core. This pattern is a hallmark of disrupted blood–brain barrier integrity and proliferative tumor angiogenesis, typically seen in high-grade gliomas [58]. The model also captures peritumoral edema, which manifests as hyperintense regions in T2-weighted or Fluid-Attenuated Inversion Recovery (FLAIR) MRI scans. This observation is clinically relevant because the presence and extent of edema correlate with the invasive potential of gliomas and often extend beyond visible contrast enhancement [59]. The Res-UNet + AG and SE variants exhibit higher fidelity in delineating these subtle edema regions compared to the baseline, which tends to under-segment infiltrative boundaries. This behavior reflects the ability of attention modules to selectively amplify weak but informative contextual features, a property that aligns with radiological expectations described in previous studies [16, 58]. Although gliomas present blurred and irregular margins, especially in infiltrative cases, the proposed models approximate their theoretical structure by reconstructing both the tumor core and surrounding edema, which is crucial for treatment planning and volume estimation.

For meningioma, the segmentation outputs maintain the tumor's well-circumscribed, rounded morphology and clear boundary contrast relative to adjacent brain parenchyma. Meningiomas typically arise from the meninges and remain extra-axial, producing high-contrast interfaces on MRI due to their encapsulated nature. The model's ability to delineate these edges sharply, as evident in Fig. A2 in Appendix, confirms that it has effectively learned the shape in priors characteristic of benign, non-infiltrative tumors. The findings correspond with previous imaging studies that meningiomas display homogeneous enhancement and well-defined borders [60]. From a computational perspective, attention-enhanced Res-UNet models likely capture these features more effectively because attention mechanisms prioritize dominant geometric cues over background textures, ensuring precise localization with minimal false-positive segmentation in surrounding tissue. The quantitative stability of meningioma segmentation further suggests that quantization does not distort feature localization in high-contrast regions, which are often less sensitive to bit-depth reduction.

In pituitary tumor segmentation, the models accurately locate lesions within the sella turcica, preserving the anatomical geometry of the pituitary gland. These tumors typically exhibit homogeneous enhancement and symmetrical shape, and deviations in segmentation contours can signify mass effects on adjacent structures such as the optic chiasm. The Res-UNet + AG variant demonstrates clear delineation of pituitary borders, maintaining consistency with clinical MRI descriptions where macroadenomas enlarge the sella without infiltrating adjacent tissue [61]. The segmentation effectively identifies structural deformation and compression effects, validating that the model captures not only local intensity differences but also spatial relationships within the confined sella region. Similar findings are reported in recent deep learning frameworks designed for pituitary segmentation, which emphasize the benefit of anatomical priors and context-aware mechanisms [61]. These results reinforce that the model's contextual awareness, facilitated by attention gating, translates to meaningful structural recognition in practice.

Overall, the segmentation outputs, as shown in Fig. A2 in Appendix, confirm that the proposed attention-quantized Res-UNet not only performs well numerically but also demonstrates anatomical plausibility and clinical coherence. Across tumor types, the model captures relevant radiological markers, ring enhancement in glioma, sharply defined boundaries in meningioma, and localized structural deformation

in pituitary tumors. Such consistency indicates that the model internalizes spatial hierarchies, which is reflective of true pathological morphology rather than relying solely on pixel intensity contrasts. Furthermore, the results imply that quantization does not compromise clinical interpretability, a critical factor for deploying AI-assisted segmentation tools in hospitals where computational resources are limited. Collectively, these findings underscore that the proposed method achieves a meaningful intersection between computational efficiency and medical validity. Hence, it supports the potential for integration into Computer-Aided Diagnosis (CADx) systems and preoperative assessment pipelines.

IV. Conclusion

In conclusion, the research successfully presents an enhanced Res-UNet-based brain tumor segmentation model that integrates multiple attention mechanisms, AG, SE Block, and CBAM, and is further optimized using DRQ. The proposed framework effectively bridges the gap between segmentation accuracy and computational efficiency, offering a practical and deployable solution for medical imaging applications. Among the evaluated variants, Res-UNet + AG demonstrates the most balanced performance, achieving a mean IoU of 0.845, with a negligible reduction of 0.0004 IoU after quantization, confirming that efficiency optimization does not significantly compromise segmentation accuracy. Additionally, quantization reduces the model size by approximately 75% (from 127.21 MB to 32.39 MB) and enhances inference speed by 37% (from 0.3143 s to 0.1973 s per image). These improvements make the model particularly suitable for real-time inference on resource-constrained devices such as medical workstations, portable diagnostic systems, and embedded edge computing platforms.

Beyond quantitative metrics, the proposed model also demonstrates strong alignment with clinical and anatomical characteristics of various brain tumor types. The segmentation results accurately delineate distinct pathological patterns: the ring enhancement and peritumoral edema of gliomas, the well-defined encapsulated structure of meningiomas, and the localized sella turcica region in pituitary tumors. Such anatomical coherence indicates that the model not only learns discriminative features but also internalizes medical imaging priors that are meaningful for clinical interpretation. This dual strength (computational efficiency and clinical relevance) reinforces the model's potential as an assistive diagnostic tool capable of enhancing radiologist workflow efficiency, supporting early detection, and improving consistency in manual segmentation tasks.

The research highlights that attention mechanisms play complementary roles: AG provides selective feature filtering for interpretability and generalization, SE Block improves gradient flow and channel-level discrimination, and CBAM enhances contextual understanding through dual-stage refinement. When combined with quantization, these mechanisms maintain high representational power even under reduced numerical precision, proving that lightweight models can still achieve robust medical segmentation outcomes. Collectively, these design strategies position the model as a strong candidate for integration into AI-driven clinical decision support systems, where accuracy, reliability, and inference speed are all critical factors.

Nevertheless, several research directions remain open. Future work should explore QAT to further minimize accuracy loss and potentially surpass the performance of post-training quantization. Expanding evaluations across multi-institutional and multi-modal datasets will improve generalizability and validate the robustness of the model in diverse imaging environments. Incorporating federated learning can also be valuable for maintaining data privacy while enabling large-scale clinical collaboration. Additionally, deploying and testing the model within actual hospital infrastructures integrated into Picture Archiving and Communication Systems (PACS) or real-time diagnostic pipelines will provide essential insights into usability, interpretability, and workflow compatibility.

Finally, while the research confirms the quantitative and qualitative validity of the proposed method through alignment with medical theories, prospective clinical validation involving expert radiologists remains a crucial next step. Such validation will ensure that segmentation decisions made by the model are trustworthy in high-stakes diagnostic contexts. Overall, the proposed Res-UNet with AG and quantization stands as a practical and scientifically grounded contribution to the development of efficient, interpretable, and clinically applicable AI systems for brain tumor segmentation. It can mark a meaningful stride toward the realization of real-world intelligent medical imaging solutions.

Author Contribution

Conceived and designed the analysis, K. A. and D. R.; Collected the data, K. A.; Contributed data or analysis tools, K. A.; Performed the analysis, K. A. and D. R.; and Wrote the paper, K. A.

Data Availability

The data that support the findings of the research are available in Kaggle at https://www.kaggle.com/datasets/nikhilroxtomar/brain-tumor-segmentation and

https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. These data were derived from the following resources available in the public domain: Figshare https://figshare.com/articles/dataset/brain_tumor_dataset/1512427, Sartaj Dataset https://www.kaggle.com/dsv/12745533, and Br35H https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no.

REFERENCES

[1] B. S. Gandhi, S. A. U. Rahman, A. Butar, and A. Victor, "Brain tumor segmentation and detection in Magnetic Resonance Imaging (MRI) using convolutional neural network," in *Brain tumor MRI image segmentation using deep learning techniques*. Elsevier, 2022, pp. 37–57.

[2] S. He, Y. Feng, P. E. Grant, and Y. Ou, "Segmentation ability map: Interpret deep features for medical image segmentation," *Medical image analysis*, vol. 84, 2023.

[3] C. M. Kumar and J. S. Sankar, "Comparative analysis of convolutional neural networks for brain tumor detection: A study of VGG16, ResNet, Inception, and DenseNet models," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. Salem, India: IEEE, June 5–7, 2024, pp. 41–46.

[4] L. Cao, J. Li, and S. Chen, "Multi-target segmentation of pancreas and pancreatic tumor based on fusion of attention mechanism," *Biomedical Signal Processing and Control*, vol. 79, 2023.

[5] P. K. Dash and D. S. Sisodia, "SEIMB-NET: A squeeze and excitation driven lightweight model for classification of brain tumors using magnetic resonance imaging," in *2024 2nd World Conference on Communication & Computing (WCONF)*. Raipur, India: IEEE, July 12–14 2024, pp. 1–6.

[6] N. Shyamala and S. Mahaboobbasha, "Convolutional block attention module-based deep learning model for MRI brain tumor identification (ResNet-CBAM)," in *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*. Trichy, India: IEEE, Sep. 18–20, 2024, pp. 1603–1608.

[7] L. Huang, A. Miron, K. Hone, and Y. Li, "Segmenting medical images: From UNet to res-UNet and nnUNet," in *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*. Guadalajara, Mexico: IEEE, June 26–28, 2024, pp. 483–489.

[8] X. Li, Z. Fang, R. Zhao, and H. Mo, "Brain tumor MRI segmentation method based on improved Res-UNet," *IEEE Journal of Radio Frequency Identification*, vol. 8, pp. 652–657, 2024.

[9] X. Fang, H. Liu, G. Xie, Y. Zhang, and D. Liu, "Deep neural network compression method based on product quantization," in *2020 39th Chinese Control Conference (CCC)*. Shenyang, China: IEEE, July 27–29, 2020, pp. 7035–7040.

[10] J. Liu, L. Niu, Z. Yuan, D. Yang, X. Wang, and W. Liu, "PD-Quant: Post-training quantization based on prediction difference metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 24 427–24 437.

[11] D. Choi, J. Park, and H. Kim, "HLQ: Hardware-friendly logarithmic quantization aware training for power-efficient low-precision CNN models," *IEEE Access*, vol. 12, pp. 159 611–159 621, 2024.

[12] Z. W. Awan, S. Khalid, and S. Gul, "A theoretical CNN compression framework for resource-restricted environments," in *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*. Rawalpindi, Pakistan: IEEE, May 24–26, 2022, pp. 1–8.

[13] M. Singh, L. Mohanty, N. Gupta, Y. Bansal, and S. Garg, "Q-Net compressor: Adaptive quantization for deep learning on resource-constrained devices," in *2024 International Conference on Computing, Sciences and Communications (ICCSC)*. Ghaziabad, India: IEEE, Oct. 24–25, 2024, pp. 1–6.

[14] J. Zhang, Z. Jiang, J. Dong, Y. Hou, and B. Liu, "Attention gate ResU-Net for automatic MRI brain tumor segmentation," *IEEE Access*, vol. 8, pp. 58 533–58 545, 2020.

[15] A. F. M. M. Rahman and M. A. Hossain, "Attention-refined U-Net with skip connections for effective brain tumor segmentation from MRI images," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. Cox's Bazar, Bangladesh: IEEE, Dec. 13–15, 2023, pp. 1–6.

[16] T. Li, J. J. Liu, Y. Tai, and Y. Tian, "Brain tumor segmentation with attention-based U-Net," in *Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering*, vol. 12079. Xi'an, China: SPIE, Oct. 8–10, 2021, pp. 147–155.

[17] Y. Z. Fang and J. D. Huang, "Enhancing brain tumor segmentation with deep supervision and attention mechanisms: Advances in the nnU-Net framework," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. Athens, Greece: IEEE, May 27–30, 2024, pp. 1–4.

[18] Z. Wang, Y. Zou, and P. X. Liu, "Hybrid dilation and attention residual U-Net for medical image segmentation," *Computers in biology and medicine*, vol. 134, 2021.

[19] X. Gan, L. Wang, Q. Chen, Y. Ge, and S. Duan, "GAU-Net: U-Net based on global attention mechanism for brain tumor segmentation," *Journal of Physics: Conference Series*, vol. 1861, no. 1, pp. 1–8, 2021.

[20] D. Zhu, "Attention-based U-Net denoising network," in *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*. Jinzhou, China: IEEE, Aug. 18–20, 2023, pp. 746–750.

[21] K. Yamamoto, "Learnable companding quantization for accurate low-bit neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, June 19–25, 2021, pp. 5029–5038.

[22] S. Dai, R. Venkatesan, M. Ren, B. Zimmer, W. Dally, and B. Khailany, "VS-Quant: Per-vector scaled quantization for accurate low-precision neural network inference," in *Proceedings of Machine Learning and Systems 3 (MLSys 2021)*, vol. 3, 2021, pp. 873–884.

[23] A. S. Molahosseini and H. Vandierendonck, "Half-precision floating-point formats for PageRank: Opportunities and challenges," in *2020 IEEE High Performance Extreme Computing Conference (HPEC)*. Waltham, MA, USA: IEEE, Sep. 22–24, 2020, pp. 1–7.

[24] Y. Jung, H. Kim, Y. Choi, and L. S. Kim, "Quantization-error-robust deep neural network for embedded accelerators," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 2, pp. 609–613, 2021.

[25] X. Wangl, Y. Zhong, and J. Dong, "A new low-bit quantization algorithm for neural networks," in *2023 42nd Chinese Control Conference (CCC)*. Tianjin, China: IEEE, July 24–26, 2023, pp. 8509–8514.

[26] D. Upadhyay, S. Malhotra, M. Gupta, and S. Mishra, "Implementation of pruned and quantized semantic segmentation neural network using Cambridge-Driving Labeled Video Database (CamVid) dataset," in *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*. Dehradun, India: IEEE, March 15–16, 2024, pp. 1–6.

[27] M. AskariHemmat, S. Honari, L. Rouhier, C. S. Perone, J. Cohen-Adad, Y. Savaria, and J. P. David, "U-Net fixed-point quantization for medical image segmentation," in *International Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*. Shenzhen, China: Springer, Oct. 13, 2019, pp. 115–124.

[28] X. Xu, Q. Lu, L. Yang, S. Hu, D. Chen, Y. Hu, and Y. Shi, "Quantization of fully convolutional networks for accurate biomedical image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, June 18–22, 2018, pp. 8300–8308.

[29] C. Qu, R. Zhao, Y. Yu, B. Liu, T. Yao, J. Zhu, B. A. Landman, Y. Tang, and Y. Huo, "Post-training quantization for 3d medical image segmentation: A practical study on real inference engines," 2025. [Online]. Available: https://arxiv.org/pdf/2501.17343

[30] R. Zhang and A. C. S. Chung, "MedQ: Lossless ultra-low-bit neural network quantization for medical image segmentation," *Medical Image Analysis*, vol. 73, 2021.

[31] C. W. Lin, Y. Hong, and J. Liu, "Aggregation-and-attention network for brain tumor segmentation," *BMC Medical Imaging*, vol. 21, pp. 1–12, 2021.

[32] R. T. Fernandes, G. R. Teixeira, E. C. Mamere, G. A. Bandeira, and A. E. Mamere, "The 2021 World Health Organization classification of gliomas: An imaging approach," *Radiologia Brasileira*, vol. 56, pp. 157–161, 2023.

[33] G. Mahesh and K. M. Yogesh, "Brain tumor detection and classification using MRI images," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 12, no. 10, pp. 856–863, 2024.

[34] K. R. Laukamp, F. Thiele, G. Shakirin, D. Zopfs, A. Faymonville, M. Timmer, D. Maintz, M. Perkuhn, and J. Borggrefe, "Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI," *European Radiology*, vol. 29, no. 1, pp. 124–132, 2019.

[35] D. Sreedhar, "Evaluating the clinical applicability of neural networks for meningioma tumor segmentation on multiparametric 3D MRI," in *2024 International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA: IEEE, Dec. 18–20, 2024, pp. 1308–1313.

[36] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain tumor detection based on deep learning approaches and magnetic resonance imaging," *Cancers*, vol. 15, no. 16, pp. 1–29, 2023.

[37] N. Tomar, "Brain tumor segmentation." [Online].

Available: https://www.kaggle.com/datasets/nikhilroxtomar/brain-tumor-segmentation

[38] M. Nickparvar, "Brain tumor MRI dataset." [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data

[39] S. F. Rabby, M. A. Arafat, and T. Hasan, "BT-Net: An end-to-end multi-task architecture for brain tumor classification, segmentation, and localization from MRI images," *Array*, vol. 22, pp. 1–14, 2024.

[40] S. Saifullah and R. Dreżewski, "Redefining brain tumor segmentation: A cutting-edge convolutional neural networks-transfer learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 3, pp. 2583–2591, 2024.

[41] N. Musthafa, M. M. Masud, and Q. Memon, "Advancing early-stage brain tumor detection with segmentation by modified_Unet," in *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, Yokohama, Japan, May 17–19, 2024, pp. 52–57.

[42] H. Alquran, M. Alslatie, A. Rababah, and W. A. Mustafa, "Improved brain tumor segmentation in MR images with a modified U-Net," *Applied Sciences*, vol. 14, no. 15, pp. 1–16, 2024.

[43] J. Wang, S. Y. Lu, S. H. Wang, and Y. D. Zhang, "RanMerFormer: Randomized vision transformer with token merging for brain tumor classification," *Neurocomputing*, vol. 573, pp. 1–12, 2024.

[44] M. M. Islam, P. Barua, M. Rahman, T. Ahammed, L. Akter, and J. Uddin, "Transfer learning architectures with fine-tuning for brain tumor classification using magnetic resonance imaging," *Healthcare Analytics*, vol. 4, pp. 1–10, 2023.

[45] S. G. De Benedictis, G. Gargano, and G. Settembre, "Enhanced MRI brain tumor detection and classification via topological data analysis and low-rank tensor decomposition," *Journal of Computational Mathematics and Data Science*, vol. 13, pp. 1–13, 2024.

[46] A. M. D. Simo, A. T. Kouanou, V. Monthe, M. K. Nana, and B. M. Lonla, "Introducing a deep learning method for brain tumor classification using MRI data towards better performance," *Informatics in Medicine Unlocked*, vol. 44, pp. 1–24, 2024.

[47] J. Peng and Y. Wang, "Medical image segmentation with limited supervision: A review of deep network models," *IEEE Access*, vol. 9, pp. 36 827–36 851, 2021.

[48] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[49] S. Mangayarkarasi, M. R. Asha, C. Kishore, M. Malathi, and V. Anitha, "Brain tumor segmentation using Res-Unet," in *2024 International Conference on Smart Electronics and Communication Systems (ISENSE)*. Kottayam, India: IEEE, Dec. 6–7, 2024, pp. 1–6.

[50] Z. Jia, H. Zhu, J. Zhu, and P. Ma, "Two-branch network for brain tumor segmentation using attention mechanism and super-resolution reconstruction," *Computers in Biology and Medicine*, vol. 157, 2023.

[51] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, pp. 1–8, 2022.

[52] Q. Xia, H. Zheng, H. Zou, D. Luo, H. Tang, L. Li, and B. Jiang, "A comprehensive review of deep learning for medical image segmentation," *Neurocomputing*, vol. 613, 2025.

[53] A. Raju and N. Sinha, "SSEGEP: Small segment emphasized performance evaluation metric for medical image segmentation," *Journal of Machine Learning in Fundamental Sciences*, vol. 2022, no. 1, pp. 1–15, 2022.

[54] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[55] X. Zhang, Y. Lin, L. Li, J. Zeng, X. Lan *et al.*, "MA-ResUNet: Multi-attention optic cup and optic disc segmentation based on improved u-net," *IET Image Processing*, vol. 18, no. 12, pp. 3128–3142, 2024.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, June 18–22, 2018, pp. 7132–7141.

[57] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, Sep. 8–14, 2018, pp. 3–19.

[58] S. Cha, "Update on brain tumor imaging: From anatomy to physiology," *American Journal of Neuroradiology*, vol. 27, no. 3, pp. 475–487, 2006.

[59] I. Blystad, J. B. M. Warntjes, Ö. Smedby,

P. Lundberg, E. M. Larsson, and A. Tisell, "Quantitative MRI using relaxometry in malignant gliomas detects contrast enhancement in peritumoral oedema," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.

[60] L. Yang, T. Wang, J. Zhang, S. Kang, S. Xu, and K. Wang, "Deep learning–based automatic segmentation of meningioma from T1-weighted contrast-enhanced MRI for preoperative meningioma differentiation using radiomic features," *BMC Medical Imaging*, vol. 24, no. 1, pp. 1–12, 2024.

[61] N. H. Lu, Y. H. Huang, K. Y. Liu, and T. B. Chen, "Deep learning-driven brain tumor classification and segmentation using non-contrast MRI," *Scientific Reports*, vol. 15, no. 1, pp. 1–24, 2025.

APPENDIX

The Appendix can be seen in the next page.

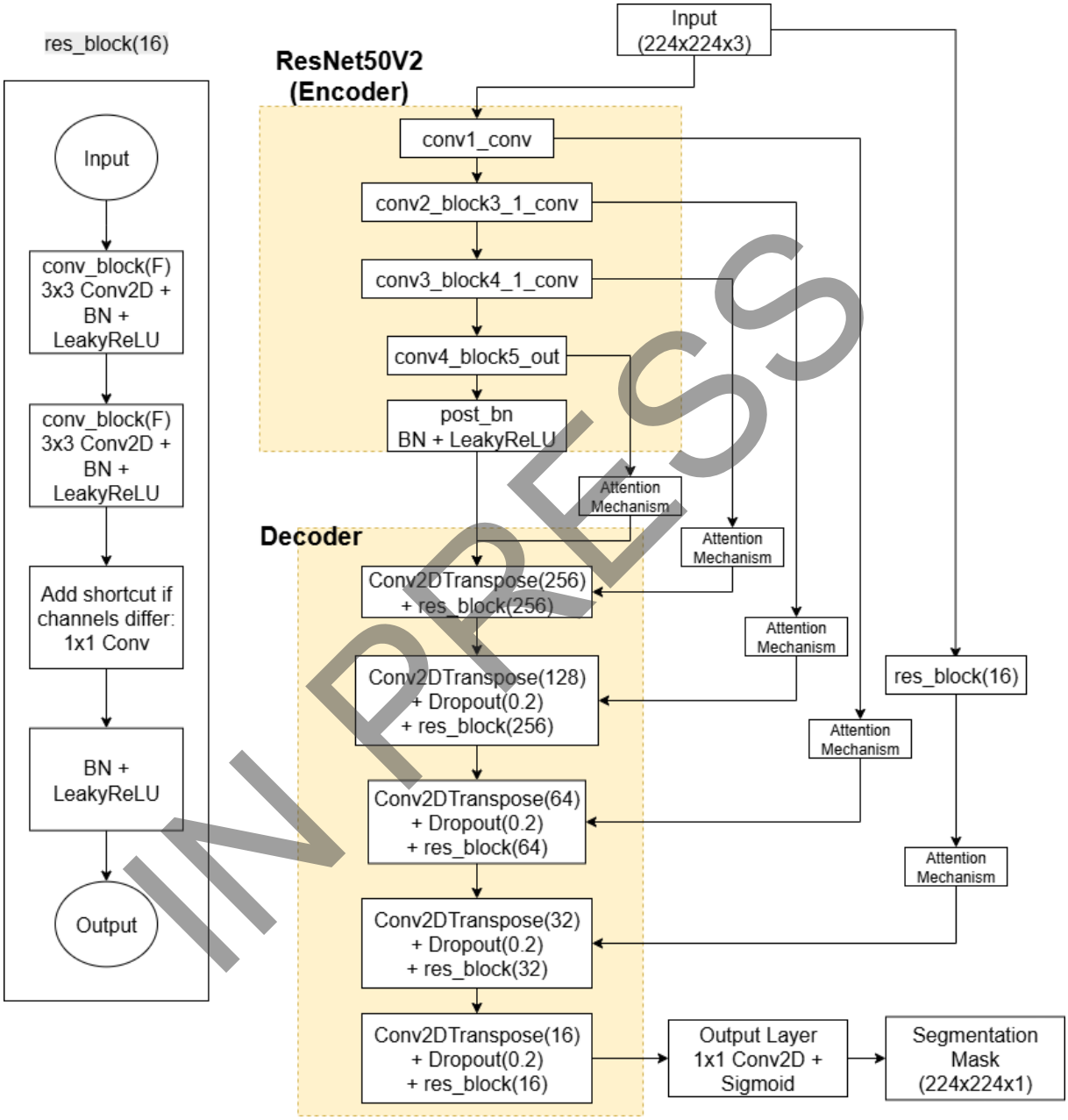

Fig. A1. Proposed architecture. Note: BatchNormalization (BN), Rectified Linear Unit (ReLU), and Convolution (conv.).

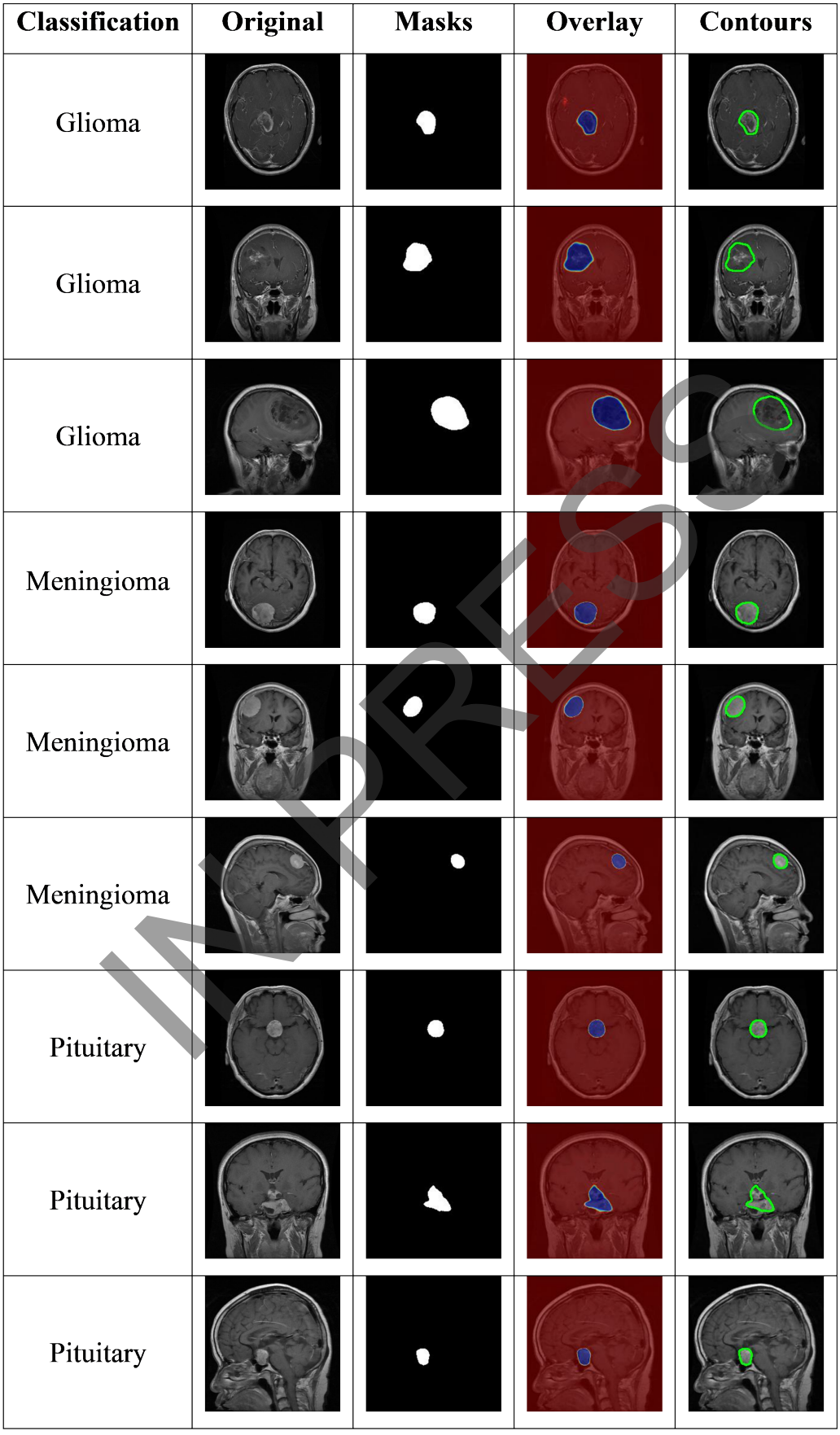

Fig. A2. Visualization of segmentation results for each model.